ROBUST PARAMETRIC CLASSIFICATION AND VARIABLE SELECTION BY A MINIMUM DISTANCE CRITERION

By Eric C. Chi* and David W. Scott†

University of California, Los Angeles and Rice University

We investigate a robust penalized logistic regression algorithm based on a minimum distance criterion. Influential outliers are often associated with the explosion of parameter vector estimates, but in the context of standard logistic regression, the bias due to outliers always causes the parameter vector to implode, that is shrink towards the zero vector. Thus, using LASSO-like penalties to perform variable selection in the presence of outliers can result in missed detections of relevant covariates. We show that by choosing a minimum distance criterion together with an Elastic Net penalty, we can simultaneously find a parsimonious model and avoid estimation implosion even in the presence of many outliers in the important small n large p situation. Implementation using an MM algorithm is described and performance evaluated.

1. Introduction. Regression, classification and variable selection problems in high dimensional data are becoming routine in fields ranging from finance to genomics. In the latter case, technologies such as expression arrays have made it possible to comprehensively query a patient's transcriptional activity at a cellular level. Patterns in these profiles can help refine subtypes of a disease according to sensitivity to treatment options or identify previously unknown genetic components of a disease's pathogenesis.

The immediate statistical challenge is finding those patterns when the number of predictors far exceeds the number of samples. To that end the Least Absolute Shrinkage and Selection Operator (LASSO) has been quite successful at addressing "the small n, big p problem" (Chen, Donoho and Saunders, 1998; Tibshirani, 1996). Indeed, ℓ_1-penalized maximum likelihood model fitting has inspired many related approaches that simultaneously do model fitting and variable selection. These approaches have been extended from linear regression to generalized linear models. In particular logistic

*Supported by DOE grant DE-FG02-97ER25308

†Supported in part by NSF grant DMS-09-07491

AMS 2000 subject classifications: Primary 62J02, 60K35; secondary 62G35

Keywords and phrases: Logistic regression, Variable selection, Robust estimation, Minimization-majorization

regression fit using the logistic deviance loss with an Elastic Net penalty (Zou and Hastie, 2005) has been well studied. Friedman, Hastie and Tibshirani (2010) provide an R package GLMNET that performs Elastic Net penalized maximum likelihood estimation for generalized linear models. Genkin, Lewis and Madigan (2007) worked out a Bayesian formulation employing the Laplace prior and applied their model to text classification. Wu et al. (2009) proposed minimizing a LASSO penalized logistic likelihood for genome wide association studies Finally, Liu et al. (2007) presented algorithms for both Elastic Net and concave Bridge penalized logistic likelihood models.

Nonetheless while ℓ_1-penalized maximum likelihood methods have proved their worth at recovering parsimonious models, noticeably less attention has been given to extending these methods to handle outliers in high dimensional data. For example in biological data, tissue samples may be mislabeled or be contaminated. Existing work centers around the Huber loss function (Owen, 2006; Rosset and Zhu, 2007). Rosset and Zhu (2007) and Wang, Zhu and Zou (2008) discuss using a Huberized hinge loss for regularized robust classification. In these references, regularized robust estimation procedures are introduced, but compelling scenarios which justify their use are not explored.

It is not surprising that outliers may bias estimation. What is less well known is that outliers can strongly influence variable selection. In this paper we identify some circumstances that motivate robust variants of penalized estimation and develop a minimum distance estimator for logistic regression. To address the $n \ll p$ scenario when predictors are correlated we add the Elastic Net penalty. We evaluate the performance of our approach through simulated and real data.

Robust methods of logistic regression are not new in the classic $n > p$ case. A broad class of solutions consists of downweighting the contribution of outlying points to the estimating equations. Downweighting can be based on extreme values in covariate space (Carroll and Pederson, 1993; Künsch, Stefanski and Carroll, 1989) or on extreme predicted probabilities (Bianco and Yohai, 1996; Carroll and Pederson, 1993; Copas, 1988).

An alternative approach is to use minimum distance estimation (Donoho and Liu, 1988). The minimum distance estimator used in this paper can also be seen as a method that downweights the contributions of outliers (Chi, 2011). The work in Bondell (2005) is similar to ours in that he considered fitting parameters by minimizing a weighted Cramér-von Mises distance. The difference between the approach proposed here and prior work is the application of regularization to handle high dimensional data and perform variable selection in the presence of outliers. Moreover, the robust loss function we propose has a particularly simple form which, when combined with

the Elastic Net penalty, can be solved very efficiently for large problems by minimizing a series of penalized least squares problems with coordinate descent.

The rest of this paper is organized as follows. We briefly explain the notation and conventions used in this paper in Section 2. In Section 3 we review maximum likelihood estimation (MLE) of the logistic regression model and demonstrate the potentially deleterious effects of outliers on variable selection with the ℓ_1-penalized MLE. We introduce our robust loss function in Section 4. In Section 5 we describe algorithms for fitting our robust logistic regression model. We describe how to select the regularization parameters in Section 6. In Sections 7 and 8 we present results on real and simulated data. Section 9 concludes with a summary of our work and also future directions.

2. Notation and conventions. In this article we will use the following notation and conventions. Vectors are denoted by boldface lowercase letters, e.g., $\mathbf{a}$. The ith entry of a vector $\mathbf{a}$ is denoted a_i. All vectors are column vectors. Matrices are denoted by boldface capital letters, e.g., $\mathbf{A}$. The element (i, j) of a matrix $\mathbf{A}$ is denoted by a_{ij}. The transpose of the ith row of a matrix $\mathbf{A}$ is $\mathbf{a}_i$. We denote the jth column of $\mathbf{A}$ with $\mathbf{a}_{(j)}$. Random variables are denoted by capital letters. The 1-norm and 2-norm of a vector $\mathbf{a}$ are denoted $\|\mathbf{a}\|_1$ and $\|\mathbf{a}\|_2$, respectively. If f is a univariate function then $f(\mathbf{a})$ and $f(\mathbf{A})$ should be interpreted as being evaluated element-wise. The kth element in a sequence is denoted by a superscript in parenthesis, e.g., $\mathbf{a}^{(k)}$ denotes the kth vector in a sequence of vectors.

Throughout this article $\mathbf{y} \in \{0, 1\}^n$ denotes a binary response vector and $\mathbf{X} \in \mathbb{R}^{n\times p}$ denotes a matrix of covariates. We will assume that the columns of $\mathbf{X}$ are centered. We utilize the generalized linear model framework of modeling the mean response as a function of affine combination of the covariates $\beta_0 + \mathbf{X}\boldsymbol{\beta}$ where $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ (McCullagh and Nelder, 1989). We will often employ the compact notations $\tilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X}) \in \mathbb{R}^{n\times(p+1)}$ and $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{p+1}$.

3. Standard logistic regression and implosion breakdown. In binary regression, we seek to predict or explain an observed response $\mathbf{y} \in \{0, 1\}^n$ using predictors $\mathbf{X} \in \mathbb{R}^{n\times p}$, where the $n \ll p$ may be expected. In typical expression microarray data we encounter $n \approx 100$ and $p \approx 10^4$, while with single nucleotide polymorphism (SNP) array data both n and p may be larger by a factor of 10. Let the conditional probabilities be given by $P(Y_i = 1|X_i = \mathbf{x}_i) = F(\tilde{\mathbf{x}}_i^\mathsf{T}\boldsymbol{\theta})$ where $F(u) = 1/(1 + \exp(-u))$. Then under this assumption, in standard logistic regression (McCullagh and Nelder, 1989) we minimize the negative log-likelihood of a linear summary of the

predictors,

$$\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{X}}\boldsymbol{\theta} - \mathbf{1}^{\mathsf{T}}\log(\mathbf{1} + \exp(\tilde{\mathbf{X}}\boldsymbol{\theta})). \tag{3.1}$$

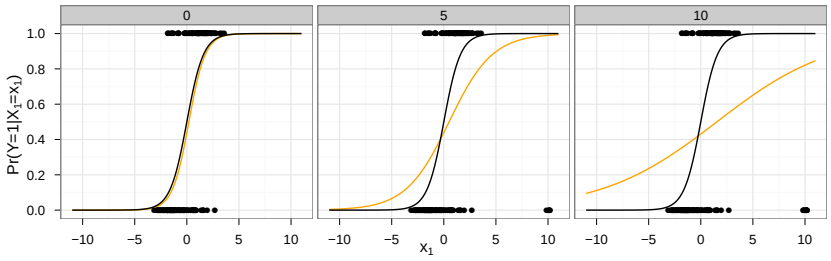
A simple univariate example illustrates the bias outliers introduce into this estimation procedure. In Figure 1(a) we see that the addition of 5 and 10 outliers among the controls shrinks $\hat{\boldsymbol{\beta}}$ towards zero. In fact, Croux, Flandre and Haesbroeck (2002) showed that with p covariates only $2p$ such outliers are required to make $\|\hat{\boldsymbol{\beta}}\|_2 < \epsilon$ for any desired ϵ. Our robust estimator, which we introduce in the next section, produces virtually the same curves shown in Figure 2(a).

The significance of this "implosion" breakdown phenomenon is that it has implications for LASSO based variable selection. Consider what happens when we add 999 noise covariates which are independent of the class labels to the scenario depicted in Figure 1(a) and perform ℓ_1-penalized logistic regression. Figure 1(b) shows the corresponding regularization paths or the values of the fitted regression coefficients as a function of the penalization parameter. As outliers are added the regularization path for the relevant covariate X_1 quickly falls into the noise.

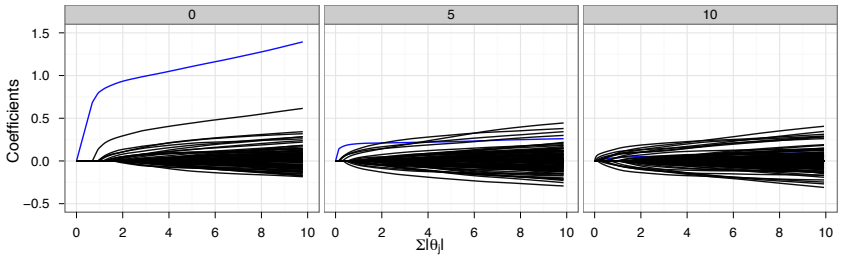
The LASSO performs continuous variable selection by shrinking regression coefficients of covariates with very low correlation with the responses to zero. If outliers are present in relevant covariates, then the combination of implosion breakdown and soft-thresholding by the LASSO can lead to missed detection of relevant covariates. In contrast in Figure 2(b) we see that the corresponding regularization paths obtained using our robust estimator are insensitive to outliers and so relevant covariates still have the chance of being selected. The message is clear; penalized robust estimation procedures have practical merit. In the next section we describe our robust estimator.

4. The Minimum Distance Estimator. Let $P_{\boldsymbol{\theta}}$ be a probability mass function (PMF), specified by a parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ for some $p \in \mathbb{N}$, believed to be generating data $Y_1, \ldots, Y_n$ that take on values in the discrete set χ. Let P be the unknown true PMF generating the data. If we actually knew the true distribution, an intuitively good solution is the one that is "closest" to the true distribution. Consequently, as an alternative to using the negative log-likelihood, we consider the L_2 distance between $P_{\boldsymbol{\theta}}$ and P. Thus, we pose the following variational optimization problem; we seek $\hat{\boldsymbol{\theta}} \in \Theta$ that minimizes

$$\sum_{y \in \chi} \left[P_{\boldsymbol{\theta}}(y) - P(y)\right]^2. \tag{4.1}$$

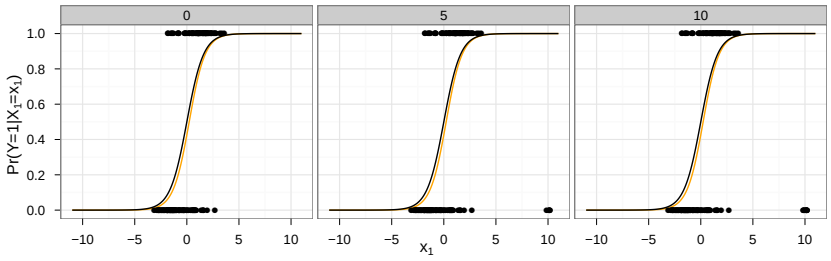


(a) Univariate regression onto X_1:The true logistic curve is black and the MLE is orange. The number of outliers (0, 5, 10) increase from the left-most to right-most panel.

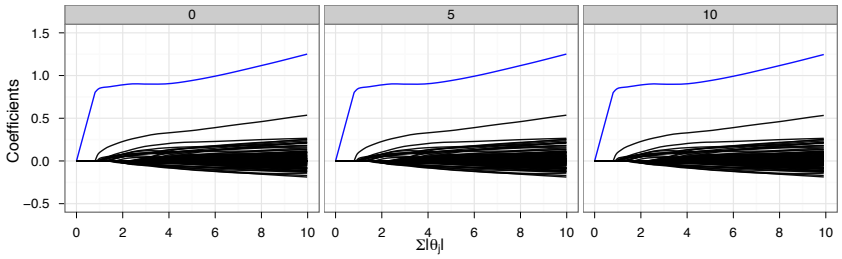


(b) Regularization paths: The path for the relevant regression coefficient β_1 is blue. The number of outliers (0, 5, 10) increase from the left-most to right-most panel; 999 irrelevant covariates have been added.

FIG 1. *MLE logistic regression with 100 cases and 100 controls.*



(a) Univariate regression onto X_1:The true logistic curve is black and the L_2E is orange. The number of outliers (0, 5, 10) increase from the left-most to right-most panel.



(b) Regularization paths: The path for the relevant regression coefficient β_1 is blue. The number of outliers (0, 5, 10) increase from the left-most to right-most panel; 999 irrelevant covariates have been added.

FIG 2. *L_2E logistic regression with 100 cases and 100 controls.*

Although finding such a $\boldsymbol{\theta}$ is impossible since P is unknown, it is possible to find a $\boldsymbol{\theta}$ that minimizes an unbiased estimate of this distance. Expanding the sum in (4.1) gives us

$$\sum_{y\in\chi} P_{\boldsymbol{\theta}}(y)^2 - 2\sum_{y\in\chi} P_{\boldsymbol{\theta}}(y)P(y) + \sum_{y\in\chi} P(y)^2.$$

The second summation is an expectation $E[P_{\boldsymbol{\theta}}(Y)]$ where Y is a random variable drawn from P. This summation can be estimated from the data by the sample mean. The third summation does not depend on $\boldsymbol{\theta}$. With these observations in mind, we use the following loss function

$$L(\boldsymbol{\theta}) = \sum_{y\in\chi} P_{\boldsymbol{\theta}}(y)^2 - \frac{2}{n}\sum_{i=1}^{n} P_{\boldsymbol{\theta}}(y_i)$$

and seek a $\hat{\boldsymbol{\theta}}$ such that $L(\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$. The estimate $\hat{\boldsymbol{\theta}}$ is called an L_2 estimate or L_2E in Scott (2001).

Note that the minimization problem is a familiar one associated with bandwidth selection for histograms and more generally for kernel density estimators (Scott, 1992). Applying a commonly used criterion in nonparametric density estimation to parametric estimation has the interesting consequence of trading off efficiency with robustness in the estimation procedure. In fact, previously Basu et al. (1998) have described a family of divergences which includes the L_2E as a special case and the MLE as a limiting case. The members of this family of divergences are indexed by a parameter that explicitly trades off efficiency for robustness. The MLE is the most efficient but least robust member in this family of estimation procedures. The L_2E represents a reasonable tradeoff between efficiency and robustness. Scott (2001, 2004) demonstrated that the L_2E has two benefits, the aforementioned robustness properties and computational tractability. The tradeoff in asymptotic efficiency is similar to that seen in comparing the mean and median as a location estimator. Indeed, while other members in this family may possess a better tradeoff, the L_2E has the advantage of admitting a simple and fast computational solution as we will show in Section 5.

We now show that the L_2E method applied to logistic regression amounts to solving a non-linear least squares problem. We seek to minimize a surrogate measure of the L_2 distance between the logistic conditional probability and the conditional probability generating the data.

If the $\mathbf{x}_i$ are unique, then we seek a $\boldsymbol{\theta}$ that "jointly" minimizes n distinct L_2 distances

$$\sum_{y\in\{0,1\}} \left[P_{\boldsymbol{\theta}}(Y_i = y|X = \mathbf{x}_i)^2 - 2y_i P_{\boldsymbol{\theta}}(Y_i = y_i|X = \mathbf{x}_i)\right] \tag{4.2}$$

where $P_{\boldsymbol{\theta}}(Y = 1|X = \mathbf{x}_i) = F(\tilde{\mathbf{X}}\boldsymbol{\theta})$. A sensible approach is to minimize a weighted average of the n distinct L_2 distances

$$\sum_{i=1}^{n} w_i \sum_{y\in\{0,1\}} \left[P_{\boldsymbol{\theta}}(Y_i = y|X = \mathbf{x}_i)^2 - 2y_i P_{\boldsymbol{\theta}}(Y_i = y_i|X = \mathbf{x}_i)\right].$$

where the weights w_i are non-negative. Up to an additive constant that does not depend on $\boldsymbol{\theta}$ the criterion in (4.2) can be compactly written as

$$\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_{\mathbf{W}}^2 + \|(\mathbf{1} - \mathbf{y}) - (\mathbf{1} - F(\tilde{\mathbf{X}}\boldsymbol{\theta}))\|_{\mathbf{W}}^2 = 2\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_{\mathbf{W}}^2$$

where $\mathbf{W} = \mathrm{diag}\{w_1, \ldots, w_n\}$ and $\|\mathbf{u}\|_{\mathbf{A}}$ denotes the semi-norm $\sqrt{\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u}}$ for $\mathbf{A}$ positive-semidefinite.

While it may be worth the effort to carefully choose the weights w_i in some cases, we will show that in the examples to follow that taking $w_i = 1/n$ for $i = 1, \ldots, n$ often works well. Thus, we will seek to minimize

$$L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) = \frac{1}{n}\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2.$$

Remarkably, minimizing this unassuming loss function produces robust logistic regression coefficients.

We note that the L_2 criterion has been used before for classification problems. Kim and Scott (2009, 2010) used the L_2 criterion to perform classification using kernel density estimates. Their application of the L_2 criterion, however, is more in line with its customary use in nonparametric density estimation whereas we use it to robustly fit a parametric model.

5. Estimation with convex quadratic majorizations. We now derive an algorithm for finding the logistic $\mathrm{L}_2\mathrm{E}$ solution by minimizing a series of convex quadratic losses. The logistic $\mathrm{L}_2\mathrm{E}$ loss is not convex. While there is no shortage of strategies for solving general unconstrained smooth non-linear minimization problems (Nocedal and Wright, 2006), we opt to minimize the $\mathrm{L}_2\mathrm{E}$ loss with a majorization-minimization (MM) algorithm (Hunter and Lange, 2004; Lange, Hunter and Yang, 2000) because it is numerically stable and easy to implement. Most importantly our MM algorithm is also easily adapted to handle LASSO-like penalties.

5.1. *Majorization-Minimization.* The strategy behind MM algorithms is to minimize a surrogate function, the majorization, instead of the original objective function. The surrogate is chosen with two goals in mind. First, an argument that decreases the surrogate should decrease the objective function. Second, the surrogate should be easier to minimize than the objective

function. Formally stated, a real-valued function h majorizes a real-valued function g at $\mathbf{v}$ if $h(\mathbf{u}) \geq g(\mathbf{u})$ for all $\mathbf{u}$ and $h(\mathbf{v}) = g(\mathbf{v})$.

In words, the surface h lies above the surface g and is tangent to g at $\mathbf{v}$. Given a procedure for constructing a majorization, we can define the MM algorithm to find a minimizer of a function g as follows. Let $\mathbf{v}^{(k)}$ denote the kth iterate: (1) find a majorization $h(\mathbf{v}; \mathbf{v}^{(k)})$ of g at $\mathbf{v}^{(k)}$; (2) set $\mathbf{v}^{(k+1)} = \arg\min_{\mathbf{v}} h(\mathbf{v}; \mathbf{v}^{(k)})$; and (3) repeat until convergence. This algorithm always takes non-increasing steps with respect to g. Consider the iteration starting at $\mathbf{v}^{(k)}$. Since $\mathbf{v}^{(k+1)}$ minimizes $h(\mathbf{v}; \mathbf{v}^{(k)})$, we have

$$g(\mathbf{v}^{(k)}) = h(\mathbf{v}^{(k)}; \mathbf{v}^{(k)}) \geq h(\mathbf{v}^{(k+1)}; \mathbf{v}^{(k)}) \geq g(\mathbf{v}^{(k+1)}).$$

By using the MM algorithm, we can convert a hard optimization problem (e.g., non-convex, non-differentiable) into a series of simpler ones (e.g., smooth convex), each of which is easier to minimize than the original.

5.2. *Majorizing the logistic L_2E loss.* Recall that the minimization problem at hand is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min}\, L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}).$$

We use the following convex quadratic majorization of $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ to find $\hat{\boldsymbol{\theta}}$.

THEOREM 5.1. *The following function majorizes $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$:*

$$L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) + \frac{2}{n}\mathbf{z}_{\tilde{\boldsymbol{\theta}}}^{\mathsf{T}}\tilde{\mathbf{X}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \frac{\eta}{n}\|\tilde{\mathbf{X}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\|_2^2, \tag{5.1}$$

where $\mathbf{z}_{\tilde{\boldsymbol{\theta}}} = 2\mathbf{G}[F(\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) - \mathbf{y}]$, $\mathbf{G} = \operatorname{diag}\{F(\tilde{\mathbf{x}}_1^{\mathsf{T}}\tilde{\boldsymbol{\theta}}), \ldots, F(\tilde{\mathbf{x}}_n^{\mathsf{T}}\tilde{\boldsymbol{\theta}})\}$, and $\eta > 0$ is sufficiently large.

Thus using the majorization (5.1) in an MM algorithm results in iterative least squares. A proof of Theorem 5.1 given in the Appendix A.1. We are able to find a simple convex quadratic majorization since the logistic L_2E loss has bounded curvature. A sharp lower bound on η is given by the maximum curvature of the logistic L_2E loss over all parameter values. The bound is derived in the Appendix. The practical implication is that the parameter η^{-1} controls the step size of our iterative solver. Consequently, in practice we set η to its lower bound to take the largest steps possible to speed up convergence.

5.3. *Iterative Least Squares.* We can express the majorization $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ (5.1) as

$$L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \eta(\tilde{\beta}_0 - \beta_0 - \frac{1}{\eta}\overline{z}_{\tilde{\boldsymbol{\theta}}})^2 + \frac{\eta}{n}\|\zeta(\tilde{\boldsymbol{\theta}}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + K(\tilde{\boldsymbol{\theta}}),$$

where

$$\begin{aligned}\overline{z}_{\tilde{\boldsymbol{\theta}}} &= n^{-1}\mathbf{1}^\mathsf{T}\mathbf{z}_{\tilde{\boldsymbol{\theta}}},\\ \zeta(\tilde{\boldsymbol{\theta}}) &= \mathbf{X}\tilde{\boldsymbol{\beta}} - \eta^{-1}(z_{\tilde{\boldsymbol{\theta}}} - \overline{z}_{\tilde{\boldsymbol{\theta}}}\mathbf{1}),\end{aligned}$$

and $K(\tilde{\boldsymbol{\theta}})$ is a constant that does not depend on $\boldsymbol{\theta}$. Since $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is separable in β_0 and $\boldsymbol{\beta}$, we can minimize $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ with respect to each of them individually. Moreover, when $\mathbf{X}$ is full rank as is often the case when $n > p$, then the solution to the normal equations is unique and the parameter updates are given by

$$\begin{aligned}\beta_0^{(m+1)} &= \beta_0^{(m)} - \eta^{-1}\overline{z}_{\boldsymbol{\theta}^{(m)}},\\ \boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} - \frac{1}{\eta}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{z}_{\boldsymbol{\theta}^{(m)}}.\end{aligned} \tag{5.2}$$

In the taxonomy of optimization procedures, the updates (5.2) constitute a Quasi-Newton method with exact line search. Note that the descent direction has a simple update since the Hessian approximation is computed only once for all iterations.

5.4. *Regularization.* The majorization given in Theorem 5.1 can be adapted for regularization. It follows immediately that $(1/2)L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) + \lambda J(\boldsymbol{\beta})$ majorizes $(1/2)L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) + \lambda J(\boldsymbol{\beta})$ for a penalty function $J : \mathbb{R}^p \to \mathbb{R}_+$ and regularization parameter $\lambda > 0$. Regularization is useful for stabilizing estimation procedures. For example if $\mathbf{X}$ is not full rank or has a large condition number we can add a ridge penalty. We then seek the minimizer to the following problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{p+1}} \frac{1}{2n}\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2 + \lambda\frac{1}{2}\|\boldsymbol{\beta}\|_2^2,$$

which we can solve by minimizing the majorization $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) + \lambda\|\boldsymbol{\beta}\|_2^2$. Since the intercept is not penalized, the intercept updates are the same as in (5.2). The update for $\boldsymbol{\beta}$ becomes

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \frac{1}{\eta}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{z}_{\boldsymbol{\theta}^{(m)}}. \tag{5.3}$$

Under suitable regularity conditions the MM algorithm for solving the ridge penalized logistic L_2E problem is guaranteed to converge to a stationary point of $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) + \lambda\|\boldsymbol{\beta}\|_2^2$. This follows from global convergence properties of MM algorithms that involve continuously differentiable objective and majorization functions (Lange, 2010). On the other hand, the MM algorithm for the unregularized version of the problem is not guaranteed to converge based on the sufficient conditions given in Lange (2010) because the objective function is not coercive (i.e., not all its level sets are compact) and the quadratic majorization is not strictly convex in $\boldsymbol{\theta}$ unless $\mathbf{X}$ is full rank. Adding the ridge penalty remedies both situations and sufficient conditions for global convergence are met. In our experience, however, when $n > p$ the iteration rules given in (5.2) do not appear to have convergence issues.

Another reason to consider regularization is to perform continuous variable selection via a LASSO-like penalty. In particular, consider the penalized majorizer for the L_2E loss regularized by the Elastic Net penalty,

$$\frac{1}{2}L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right). \tag{5.4}$$

Since our work is motivated by genomic data which is known to have correlated covariates, we will focus on the Elastic Net penalty because it produces sparse models but includes and excludes groups of correlated variables (Zou and Hastie, 2005). The LASSO, in contrast, tends to select one covariate among a group correlated covariates and exclude the rest. If groupings among the covariates are known in advance, a group LASSO penalty could be used (Yuan and Lin, 2006). The Elastic Net penalty is useful in that it performs group selection without prespecification of the groups.

We are interested in generating MM iterates $\boldsymbol{\theta}^{(m)} = \left(\beta_0^{(m)}, \boldsymbol{\beta}^{(m)}\right)$ where

$$\begin{aligned} \beta_0^{(m+1)} &= \beta_0^{(m)} - \eta^{-1}\overline{z}_{\boldsymbol{\theta}^{(m)}} \\ \boldsymbol{\beta}^{(m+1)} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{\eta}{2n}\|\zeta(\boldsymbol{\theta}^{(m)}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right). \end{aligned} \tag{5.5}$$

Before discussing how to practically solve the surrogate minimization problem, note that regardless of how (5.5) is solved, we have the following guarantee on the convergence of the MM iterates.

THEOREM 5.2. *Under suitable regularity conditions for any starting point* $\boldsymbol{\theta}^{(0)}$*, the sequence of iterates* $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ *generated by (5.5) converges to a*

stationary point of

$$\frac{1}{2n}\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right),$$

where $\lambda > 0$ *and* $\alpha \in [0, 1)$.

A proof is given in the Appendix A.2 and relies on a recent extension of the global convergence properties of MM algorithms for locally Lipschitz continuous objective and majorization functions (Schifano, Strawderman and Wells, 2010). Note that Theorem 5.2 restricts $\alpha < 1$, i.e., algorithmic convergence of the LASSO regularized logistic L_2E is not guaranteed. This condition is imposed to ensure that the majorization is strictly convex in $\boldsymbol{\beta}$. Again in our experience, the LASSO regularized logistic L_2E does not have algorithmic convergence issues.

As a final remark on algorithmic convergence note that since the ridge penalty is a special case of the Elastic Net, Theorem 5.2 implies that ridge penalized logistic L_2E (5.3) will also converge.

5.5. *Coordinate Descent.* The optimization problem for updating $\boldsymbol{\beta}$ in (5.5) can be written as an ℓ_1-penalized least squares problem.

$$\boldsymbol{\beta}^{(m+1)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, \frac{\eta}{2n}\left\|\begin{pmatrix}\boldsymbol{\zeta}(\boldsymbol{\theta}^{(m)})\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda(1-\alpha)}\mathbf{I}\end{pmatrix}\boldsymbol{\beta}\right\|_2^2 + \lambda\alpha\|\boldsymbol{\beta}\|_1.$$

Thus, minimizing the Elastic Net penalized logistic L_2E loss can be cast into a series of LASSO regressions. We choose to solve (5.5) with coordinate descent which has been shown to efficiently solve penalized regression problems when selecting relatively few groups of correlated predictors (Friedman et al., 2007; Wu and Lange, 2008).

Coordinate descent is a special case of block relaxation optimization where, in a round-robin fashion, we optimize the objective function with respect to each coordinate at a time while holding all other coordinates fixed. Formally, if we are minimizing a multivariate function f at the kth round of coordinate descent for the jth coordinate we solve

$$\theta_j^{(k)} \in \underset{\theta}{\arg\min}\, f(\theta_1^{(k)}, \ldots, \theta_{j-1}^{(k)}, \theta, \theta_{j+1}^{(k-1)}, \ldots, \theta_p^{(k-1)}). \tag{5.6}$$

The jth coordinate update during the kth round of cyclic coordinate descent of the mth MM iteration is well defined, i.e., there exists a unique

Algorithm 1 ITERATIVE L_2E SOLVER

$\boldsymbol{\theta} \leftarrow$ initial guess
repeat
  $\mathbf{p} \leftarrow F(\tilde{\mathbf{X}}\boldsymbol{\theta})$
  $\mathbf{G} \leftarrow \operatorname{diag}\{\mathbf{p} * (\mathbf{1} - \mathbf{p})\}$
  $\mathbf{z} \leftarrow 2\mathbf{G}(\mathbf{p} - \mathbf{y})$
  $\boldsymbol{\zeta} \leftarrow \mathbf{X}\boldsymbol{\beta} - \frac{1}{\eta}(\mathbf{z} - \overline{z}\mathbf{1})$
  $\beta_0 \leftarrow \beta_0 - \eta^{-1}\overline{z}$
  repeat
    for $k = 1..p$ **do**
      $\mathbf{r} \leftarrow \boldsymbol{\zeta} - (\mathbf{X}\boldsymbol{\beta} - \beta_k \mathbf{x}_k)$
      $\beta_k \leftarrow S\left(\frac{\eta}{n}\mathbf{x}_k^{\mathsf{T}}\mathbf{r}, \lambda\alpha\right) / \left[\frac{\eta}{n}\|\mathbf{x}_k\|_2^2 + \lambda(1-\alpha)\right]$
    end for
  until convergence
until convergence
return $\boldsymbol{\theta}$

minimizer in (5.6). The update $\beta_j^{(m,k)}$ has a simple form (Donoho and Johnstone, 1995) and is given by the subgradient equations to be

$$\beta_j^{(m,k)} = \frac{S\left(\frac{\eta}{n}\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{r}^{(m,k,j)}, \lambda\alpha\right)}{\frac{\eta}{n}\|\mathbf{x}_{(j)}\|_2^2 + \lambda(1-\alpha)},$$

where $\mathbf{r}^{(m,k,j)}$ is a vector of partial residuals with ith entry

$$r_i^{(m,k,j)} = \zeta_i(\boldsymbol{\theta}^{(m)}) - \left(\sum_{j'=1}^{j-1} x_{ij'}\beta_{j'}^{(m,k)} + \sum_{j'=j+1}^{p} x_{ij'}\beta_{j'}^{(m,k-1)}\right),$$

and S is the soft-threshold function

$$S(a, \lambda) = \operatorname{sign}(a)\max(|a| - \lambda, 0).$$

Algorithm 1 gives pseudocode for the resulting iterative solver. The symbol $*$ denotes the Hadamard element-wise product. In practice we also use active sets to speed up computations. That is, for a given initial $\boldsymbol{\beta}$, we only update the non-zero coordinates of $\boldsymbol{\beta}$, the active set, until there is little change in the active set parameter estimates. The non-active set parameter estimates are then updated once. If they remain zero, the Karush-Kuhn-Tucker (KKT) conditions have been met and a global minimum of (5.5) has been found. If not, then the active set is expanded to include the coordinates whose KKT conditions have been violated and the process is repeated.

6. Choosing the penalty parameters.

6.1. *Warm Starts and Calculating Regularization Paths.* We will need to compare the regression coefficients obtained at many values of the penalty parameter λ to perform model selection. Typically we can rapidly calculate regression coefficients for a decreasing sequence of values of λ through warm starts. For λ sufficiently large, only the intercept term θ_0 will come into the model. The smallest λ^* such that all regression coefficients are shrunk to zero is given by

$$\lambda^* = \frac{2}{n\alpha}\overline{y}(1-\overline{y}) \max_{j=1,\ldots,p} |\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{y}|. \tag{6.1}$$

We compute a grid of λ values equally spaced on a log scale between $\lambda_{\max} = \lambda^*$ and $\lambda_{\min} = \epsilon\lambda_{\max}$ where $\epsilon < 1$. In practice, we have found the choice of $\epsilon = 0.05$ to be useful. In general, we are not interested in making λ so small as to include all variables.

Moreover, due to the possible multi-modality of the L_2E loss, we recommend computing the regulation paths starting from a smaller regularization parameter and increasing the parameter value until $\lambda_{\max}$. Since we face multi-modality initial starting points can make a significant difference in the answers obtained.

6.2. *The heuristic for choosing starting values.* Since the logistic L_2E loss is not convex, it may have multiple local minima. For the purely lasso-penalized problem, the KKT condition at a local minimum is

$$|\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{G}(\mathbf{y} - F(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}))| \leq \lambda.$$

Equality is met whenever $\beta_j \neq 0$. Thus, the largest values of

$$|\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{G}(\mathbf{y} - F(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}))|$$

will correspond to a set of covariates which include covariates with non-zero regression coefficients. The leap of faith is that the largest values of

$$|\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{G}(\mathbf{y} - F(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}))|$$

evaluated at the null model will also correspond to a set of covariates which include covariates with non-zero regression coefficients. This idea has been used in a "swindle" rule (Wu et al., 2009), SAFE rules for discarding variables (El Ghaoui, Viallon and Rabbani, 2010), and STRONG rules for discarding variables (Tibshirani et al., 2011). In those instances the goal is to solve a smaller optimization problem. In contrast, we initialize starting parameter entries to zero rather than excluding variables with low scores from the optimization problem.

Calculate the following scores z_j

$$z_j = |\mathbf{x}_{(j)}^{\mathsf{T}} \mathbf{G}_0 (\mathbf{y} - p\mathbf{1}))|,$$

where $p = \overline{y}$ the sample mean of $\mathbf{y}$ and $\mathbf{G}_0 = p(1-p)\mathbf{I}$.

Let $\mathcal{S} = \{j : z_j \text{ is "large"}\}$. Assign the starting parameter value as follows

$$\beta_0^{(0)} = \log(\overline{y}/(1-\overline{y}))$$
$$\beta_j^{(0)} = \begin{cases} 1 & j \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

6.3. *Robust Cross-Validation.* Once we have a set of models computed at different regularization parameter values, we select the model that is optimal with respect to some criterion. We use the following robust 10-fold cross-validation scheme to select the model. After partitioning the data into 10 training and test sets, for each $i = 1, \ldots, 10$ folds we compute regression coefficients $\hat{\boldsymbol{\theta}}^{-i}(\lambda)$ for a sequence of λ's between $\lambda_{\max}$ and $\lambda_{\min}$ holding out the ith test set $\mathcal{S}_i$.

Next we refit the model using the reduced variable set $\mathcal{S}_i^c$, those with nonzero regression coefficients, and refit using logistic L_2E with $\alpha = 0$. This refitting produces less biased estimates. We are adopting the same strategy as LARS-OLS in Efron et al. (2004). Our framework, however, could adopt a more sophisticated strategy along the lines of the Relaxed LASSO in Meinshausen (2007). Henceforth let $\hat{\boldsymbol{\theta}}^{-i}(\lambda)$ denote the regression coefficients obtained after the second step. Let $d_j^{-i}(\lambda)$ denote the contribution of observation j to the L_2E loss under the model $\hat{\boldsymbol{\theta}}^{-i}(\lambda)$, i.e.,

$$d_j^{-i}(\lambda) = \left(y_j - F(\tilde{\mathbf{x}}_j^{\mathsf{T}} \hat{\boldsymbol{\theta}}^{-i}(\lambda))\right)^2.$$

We use the following criterion to choose λ^*:

$$\lambda^* = \arg\min_{\lambda} \ \text{median}_{i=1,\ldots,10} \, \text{median}_{j \in \mathcal{S}_i} \, d_j^{-i}(\lambda).$$

The reason for choosing λ in this way is due to a feature of the robust fitting procedure. Good robust models will assign unusually large values of $d_j^{-i}(\lambda)$ to outliers. Thus, the total L_2E loss is an inappropriate measure of the prediction error if influential outliers were present. On the other hand, taking the median, for example, would provide a more unbiased measure of the prediction error regardless of outliers. The final model selected would be the one that minimizes the robust prediction error criterion.

7. Simulations. In this section we report on three simulations comparing the MLE and L_2E results. The first two simulations examine the accuracy of estimation. We then follow with a simulation experiment designed to examine the variable selection properties. For the first two simulations we generated 1000 data sets, with 200 binary outcomes each associated with 4 covariates, from the logistic model specified by the likelihood in (3.1) with parameters $\beta_0 = 0$ and $\boldsymbol{\beta} = (1, 0.5, 1, 2)^\mathsf{T}$. The covariates $\mathbf{x}_i$ were drawn from one of two populations.

$$\mathbf{x}_i \overset{i.i.d.}{\sim} \begin{cases} N(\boldsymbol{\mu}, 0.16\,\mathbf{I}_p) & i = 1, \ldots, 100 \\ N(-\boldsymbol{\mu}, 0.16\,\mathbf{I}_p) & i = 101, \ldots, 200 \end{cases}$$

where $p = 4$ and $\boldsymbol{\mu} = (0.25, 0.25, 0.25, 0.25)^\mathsf{T}$. The responses were generated from the model

$$\mathbf{y}_i \overset{i.i.d.}{\sim} \textsc{Bernoulli}(F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}), 1).$$

7.1. *Estimation in Low Dimensions.* In the first scenario, we added a single outlier, $(y_{201}, \mathbf{x}_{201})$ where $y_{201} = 0$ and $\mathbf{x}_{201} = (\delta, \delta, \delta, \delta)^\mathsf{T}$ and δ took on values in $\{-0.25, 1.5, 3, 6, 12, 24\}$. In words, the 201st point was moved in covariate space along the line that runs through the centroids of the two subpopulations. In the second scenario, we added a variable number of outliers at a single location: $\{(y_i, \mathbf{x}_i)\}_{i=201}^N$, where $y_i = 0$ and $\mathbf{x}_i = (3, 3, 3, 3)^\mathsf{T}$ for $i = 201, \ldots, N$ and the number of outliers is $N = 0, 1, 5, 10, 15, 20$. For each sequence of scenarios described, we performed logistic regression and L_2E regression. Tables 1 and 2 show the mean and standard deviation for the fitted coefficient values in the first and second scenario, respectively.

The results show two features of the L_2E versus the MLE. Consider the first scenario. The MLE becomes increasingly biased towards zero as the 201st point is moved from -0.25 to to 24. In contrast, the L_2E is insensitive to the placement of the 201st point. Figure 3 shows how $\|\hat{\boldsymbol{\beta}}\|_2$ under each estimation procedure varies with the position of outlier is moved. We see that MLE values demonstrate "implosion" breakdown, i.e., $\|\hat{\boldsymbol{\beta}}\|_2$ tends towards 0 as the leverage of the 201st point increases. The L_2E estimates do not. The second observation is the cost of the L_2E's unbiasedness is increased variance as seen in the increased standard error in Table 1. The L_2E's sample standard error is greater than the MLE's for all locations of the outlier. Similar behavior is observed in the second scenario. Figure 4 shows that "implosion" breakdown ensues as outliers are added.

7.2. *Variable Selection in High Dimensions.* In the variable selection experiment we considered a high dimensional variation on the first scenario.

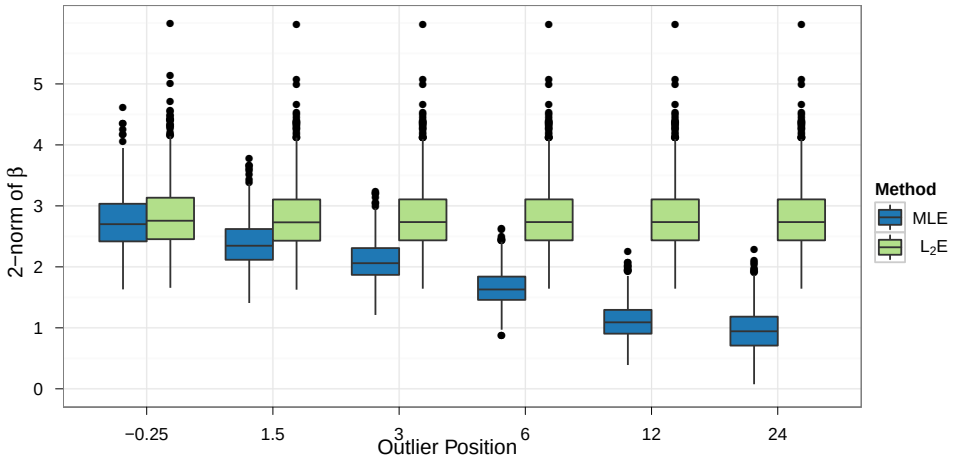


FIG 3. *The 2-norm of the regression coefficients (intercept not included) as a function of a single outliers position.*

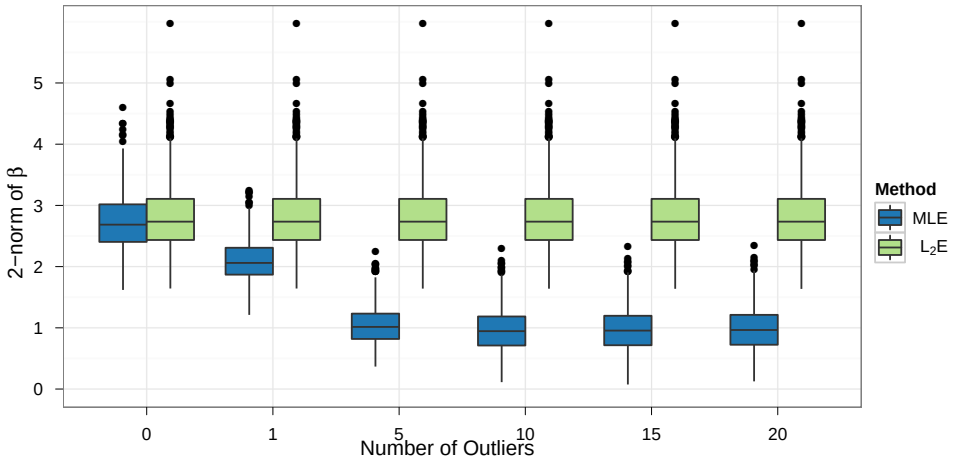


FIG 4. *The 2-norm of the regression coefficients (intercept not included) as a function of the number of outliers at a fixed position.*

TABLE 1

Varying the location of a single outlier: The estimates calculated by L_2E are essentially unbiased regardless of the location of the outlier. In contrast, the MLE results become very biased as the outlier position ranges from -0.25 to 24. The unbiasedness of the L_2E comes at a price of increased variance. The sample standard error of the L_2E is greater than that of the MLE for all outlier positions.

			MLE		L_2E	
Outlier Position	Coefficient	True Value	mean	std	mean	std
-0.25	β_0	0	-0.002	0.182	-0.005	0.192
	β_1	1	1.032	0.434	1.063	0.480
	β_2	0.5	0.526	0.424	0.539	0.463
	β_3	1	1.047	0.439	1.079	0.482
	β_4	2	2.110	0.487	2.181	0.572
1.5	β_0	0	-0.024	0.168	0.002	0.192
	β_1	1	0.868	0.394	1.052	0.476
	β_2	0.5	0.401	0.391	0.532	0.460
	β_3	1	0.880	0.396	1.068	0.478
	β_4	2	1.860	0.430	2.160	0.567
3	β_0	0	-0.022	0.157	0.002	0.192
	β_1	1	0.732	0.368	1.054	0.476
	β_2	0.5	0.296	0.369	0.533	0.460
	β_3	1	0.743	0.368	1.069	0.478
	β_4	2	1.662	0.392	2.163	0.567
6	β_0	0	-0.020	0.142	0.002	0.192
	β_1	1	0.508	0.337	1.054	0.476
	β_2	0.5	0.112	0.344	0.533	0.460
	β_3	1	0.516	0.334	1.069	0.478
	β_4	2	1.350	0.347	2.163	0.567
12	β_0	0	-0.018	0.128	0.002	0.192
	β_1	1	0.153	0.325	1.054	0.476
	β_2	0.5	-0.201	0.336	0.533	0.460
	β_3	1	0.158	0.316	1.069	0.478
	β_4	2	0.906	0.317	2.163	0.567
24	β_0	0	-0.011	0.124	0.002	0.192
	β_1	1	-0.088	0.330	1.054	0.476
	β_2	0.5	-0.431	0.331	0.533	0.460
	β_3	1	-0.086	0.315	1.069	0.478
	β_4	2	0.641	0.324	2.163	0.567

We generated 10 data sets each with $n = 500$ observations. The covariates

TABLE 2

Varying the number of outliers at a fixed location: The estimates calculated by L_2E are essentially unbiased of the number of outliers. In contrast, the MLE results become very biased as outliers are added. The unbiasedness of the L_2E comes at a price of increased variance. The sample standard error of the L_2E is greater than that of the MLE for all numbers of outliers.

			MLE		L_2E	
Number of Outliers	Coefficient	True Value	mean	std	mean	std
0	$\boldsymbol{\beta}_0$	0	0.0049	0.1824	0.0021	0.1923
	$\boldsymbol{\beta}_1$	1	1.0258	0.4326	1.0537	0.4759
	$\boldsymbol{\beta}_2$	0.5	0.5213	0.4225	0.5327	0.4599
	$\boldsymbol{\beta}_3$	1	1.0405	0.4376	1.0690	0.4782
	$\boldsymbol{\beta}_4$	2	2.0994	0.4853	2.1630	0.5666
1	$\boldsymbol{\beta}_0$	0	-0.0221	0.1573	0.0021	0.1923
	$\boldsymbol{\beta}_1$	1	0.7324	0.3679	1.0537	0.4759
	$\boldsymbol{\beta}_2$	0.5	0.2956	0.3690	0.5327	0.4599
	$\boldsymbol{\beta}_3$	1	0.7431	0.3681	1.0690	0.4782
	$\boldsymbol{\beta}_4$	2	1.6620	0.3924	2.1629	0.5666
5	$\boldsymbol{\beta}_0$	0	-0.0898	0.1258	0.0021	0.1923
	$\boldsymbol{\beta}_1$	1	0.0864	0.3201	1.0537	0.4759
	$\boldsymbol{\beta}_2$	0.5	-0.2628	0.3272	0.5327	0.4599
	$\boldsymbol{\beta}_3$	1	0.0905	0.3082	1.0690	0.4782
	$\boldsymbol{\beta}_4$	2	0.8300	0.3125	2.1629	0.5666
10	$\boldsymbol{\beta}_0$	0	-0.1101	0.1237	0.0021	0.1923
	$\boldsymbol{\beta}_1$	1	-0.0735	0.3296	1.0536	0.4759
	$\boldsymbol{\beta}_2$	0.5	-0.4167	0.3332	0.5326	0.4599
	$\boldsymbol{\beta}_3$	1	-0.0709	0.3153	1.0690	0.4782
	$\boldsymbol{\beta}_4$	2	0.6586	0.3226	2.1628	0.5667
15	$\boldsymbol{\beta}_0$	0	-0.1172	0.1237	0.0021	0.1923
	$\boldsymbol{\beta}_1$	1	-0.1268	0.3354	1.0536	0.4759
	$\boldsymbol{\beta}_2$	0.5	-0.4696	0.3385	0.5326	0.4599
	$\boldsymbol{\beta}_3$	1	-0.1245	0.3208	1.0689	0.4782
	$\boldsymbol{\beta}_4$	2	0.6048	0.3282	2.1627	0.5667
20	$\boldsymbol{\beta}_0$	0	-0.1216	0.1238	0.0021	0.1923
	$\boldsymbol{\beta}_1$	1	-0.1586	0.3393	1.0535	0.4759
	$\boldsymbol{\beta}_2$	0.5	-0.5016	0.3423	0.5326	0.4599
	$\boldsymbol{\beta}_3$	1	-0.1566	0.3246	1.0689	0.4782
	$\boldsymbol{\beta}_4$	2	0.5735	0.3318	2.1626	0.5668

were drawn from one of three multivariate normal populations

$$\mathbf{x}_i \overset{i.i.d.}{\sim} \begin{cases} N(\boldsymbol{\mu}, 0.75\,\mathbf{I}_p) & i = 1, \ldots, 200 \\ N(-\boldsymbol{\mu}, 0.75\,\mathbf{I}_p) & i = 201, \ldots, 400 \\ N(\boldsymbol{\nu}, 0.25\,\mathbf{I}_p) & i = 401, \ldots, 500 \end{cases}$$

where $p = 500$ and

$$\boldsymbol{\mu} = (\underbrace{0.3, \ldots, 0.3}_{50}, \underbrace{0, \ldots, 0}_{450})$$
$$\boldsymbol{\nu} = (\underbrace{1, \ldots, 1}_{50}, \underbrace{0, \ldots, 0}_{450}).$$

The responses were generated as follows

$$\mathbf{y}_i \overset{i.i.d.}{\sim} \begin{cases} \text{BERNOULLI}(F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}), 1) & i = 1, \ldots, 400 \\ 0 & i = 401, \ldots, 500, \end{cases}$$

where $\beta_0 = 0$ and

$$\boldsymbol{\beta} = (\underbrace{1, \ldots, 1}_{50}, \underbrace{0, \ldots, 0}_{450}) \in \mathbb{R}^{500}.$$

We then performed Elastic Net penalized regression ($\alpha = 0.6$) with the MLE and L_2E. For L_2E, we chose the initial starting point according to the heuristic described in Section 6.2. To perform model selection we generated regularization paths. That is we calculated penalized regression coefficients for a range of λ values using the robust cross-validation method described in Section 6. To perform the elastic net penalized logistic regression we used the **glmnet** package in R (Friedman, Hastie and Tibshirani, 2010). We also compared the robust classifier of Wang, Zhu and Zou (2008) - the Hybrid Huberized Support Vector Machine (HHSVM) using an MM algorithm. Details of the implementation can be found in a supplement on the author's website.[1] Wang, Zhu and Zou (2008) provide code for computing the solutions paths of the HHSVM, but the algorithm used calculates the paths for a varying LASSO regularization parameter with a fixed ridge regularization parameter because they can be computed quickly by exploiting the piecewise linearity of the paths under that parameterization of the Elastic Net. Our implementation calculates regularization paths using the Elastic Net parameterization used in this manuscript.

Tables 3 and Table 4 show the number of true positives and false positives respectively for each method. We see that in scenarios of heavy contamination the L_2E demonstrates superior sensitivity and specificity compared to both the MLE and HHSVM. It is interesting to note that the MLE tends to be more sensitive than the HHSVM, but at a cost of being drastically less specific.

[1] http://www.stat.lsa.umich.edu/ jizhu/code/hhsvm/

TABLE 3
True positive count with $n = p = 500$ and 50 true covariates. L_2E is the most sensitive method. HHSVM is the least sensitive method.

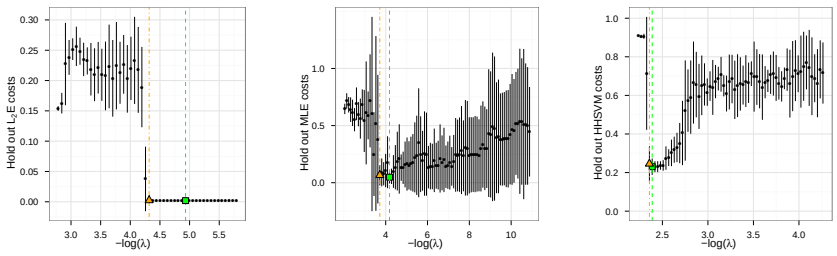
	Replicate									
	1	2	3	4	5	6	7	8	9	10
MLE	14	10	8	10	1	10	0	14	11	15
HHSVM	1	3	2	2	1	2	1	2	4	2
L_2E	48	47	48	49	48	48	49	46	48	49

TABLE 4
False positive count with $n = p = 500$ and 50 true covariates. L_2E is the most specific method. MLE is the least specific method.

	Replicate									
	1	2	3	4	5	6	7	8	9	10
MLE	141	95	56	148	0	141	0	128	136	170
HHSVM	0	4	1	1	1	0	1	0	0	0
L_2E	0	0	2	0	0	0	1	1	0	1

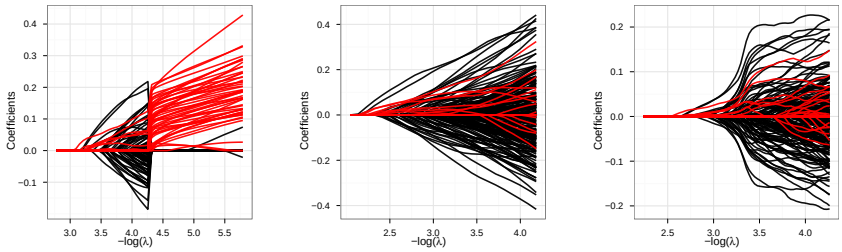
Figure 5 shows the robust cross validation curves for the three methods for one of the replicates. Note the large jump in the L_2E curve. By choosing the starting L_2E point by our heuristic, a local minimum different from the MLE solution is found. For sufficiently large λ, however, the local minimum vanishes, and the regularization paths mimic the MLE regularization paths.

8. Real data examples.



(a) L_2E (b) MLE (c) HHSVM

FIG 5. *Robust 10-fold cross-validation curves for the three methods. The green square indicates the minimizing λ. The orange triangle indicates the 1-MAD rule λ.*



(a) L_2E (b) MLE (c) HHSVM

FIG 6. *Regularization paths for the three methods. True regression coefficients for true covariates are in red.*

8.1. *An $n > p$ example: Predicting abnormal and normal vertebral columns.* We first consider a real data set in the $n > p$ regime. We present results on the vertebral column data set from the UCI machine learning repository, as described by Frank and Asuncion (2010). The data set consists of 310 patients which have been classified as belonging to one of three groups: Normal (100 patients), Disk Hernia (60 patients), Spondylolisthesis (150 patients). In addition to a classification label, six predictor variables are recorded for each patient: pelvic incidence (PI), pelvic tilt (PT), lumbar lordosis angle (LLA), sacral slope (SS), pelvic radius (PR) and grade of spondylolisthesis (GS). All six predictor variables are continuous valued.

We consider the two class problem of discriminating normal vertebral columns from abnormal ones (Disk Hernia and Spondylolisthesis). Figure 7 plots the values of individual covariates for each patient. Table 5 shows the correlations between pairs of attributes. Note that the attributes for Disk Hernia and Normal patients overlap a good deal. We may expect similar results as seen in the second simulation scenario described in Section 7.1 where Disk Hernia patients play the role of a cluster of outlying observations. Due to the correlation, however, the outlying observations are not as distinctly outlying as seen in the simulation examples of Section 7.1. Consequently, it also might be anticipated that there will not be differences between the MLE and L_2E regularization paths. Indeed, Figure 8 shows the resulting regularization paths generated by the MLE and logistic L_2E for $\alpha = 0.2$. The paths are very similar for both methods for other values of α and are not shown. Different initial starting points did not change the resulting logistic L_2E regularization paths.

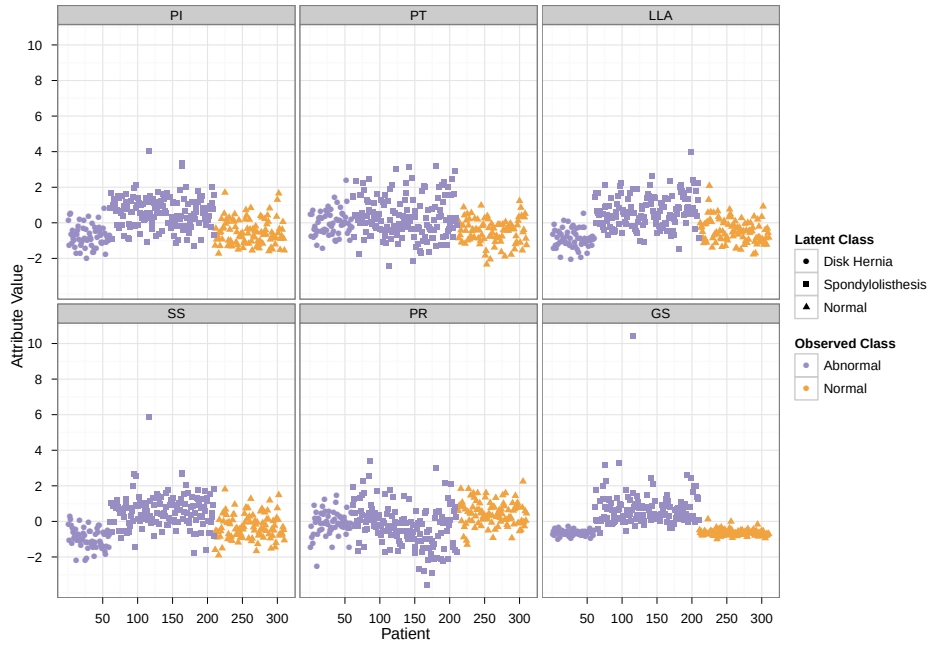


FIG 7. *Six biomechanical attributes: pelvic incidence (PI), pelvic tilt (PT), lumbar lordosis angle (LLA), sacral slope (SS), pelvic radius (PR) and grade of spondylolisthesis (GS). There are three true classes of observations (Disk Hernia, Spondylolisthesis, and Normal). Disk Hernia and Spondylolisthesis are lumped into the class Abnormal.*

	PI	PT	LLA	SS	PR	GS
PI	1.00	0.63	0.72	0.81	-0.25	0.64
PT	0.63	1.00	0.43	0.06	0.03	0.40
LLA	0.72	0.43	1.00	0.60	-0.08	0.53
SS	0.81	0.06	0.60	1.00	-0.34	0.52
PR	-0.25	0.03	-0.08	-0.34	1.00	-0.03
GS	0.64	0.40	0.53	0.52	-0.03	1.00

TABLE 5

Correlations among the six biomechanical attributes in the vertebrae data set.

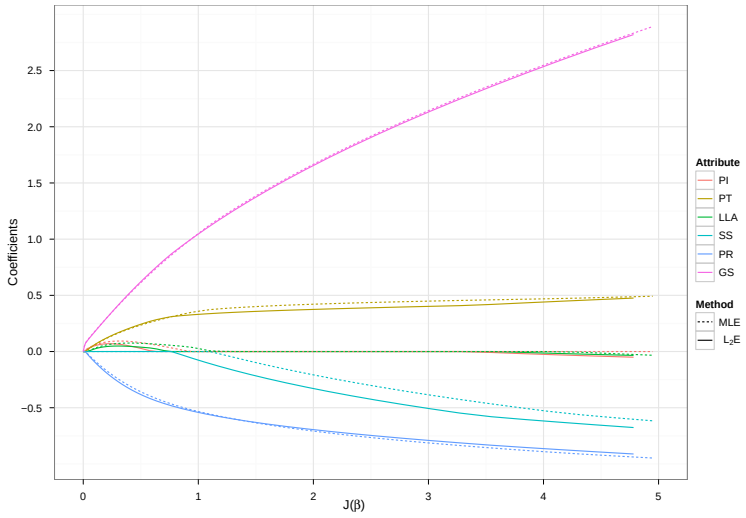


FIG 8. *The regularization ($\alpha = 0.2$) paths for the MLE and L_2E are very similar for the six biomechanical attributes in the vertebrae data set.*

8.2. *An $n \ll p$ example: a genome wide association study.* We examine the lung cancer data of Amos et al. (2008). The purpose of this genome wide association study was to identify risk variants for lung cancer. The authors employed a two stage study using 315,450 tagging SNPs in 1,154 current and former (ever) smokers of European ancestry and 1,137 frequency matched, ever-smoking controls from Houston, Texas in the discovery stage. The most significant SNPs found in the discovery phases were then tested in a larger replication set. Two SNPs, rs1051730 and rs8034191, on chromosome 15 were found to be significantly associated with lung cancer risk in the validation set.

In this section we reexamine the discovery data using logistic L_2E and the logistic MLE. Note that it is current practice of geneticists to do univariate

inference with an adjustment for multiple testing and this approach was taken in Amos et al. (2008). Taking a multivariate approach as will be done in this section, however, allows the analyst to take into account dependencies between the SNPs. As an initial comparison we consider a subset of the entire data set and restrict our analysis to SNPs on chromosome 15. We impute missing genotypes at a SNP by using the MACH 1.0 package, a Markov Chain based haplotyper (Li, Ding and Abecasis, 2006). After missing data are imputed and keeping only imputations with a quality score of at least 0.9, 8,701 SNPs are retained on 1152 cases and 1136 controls.

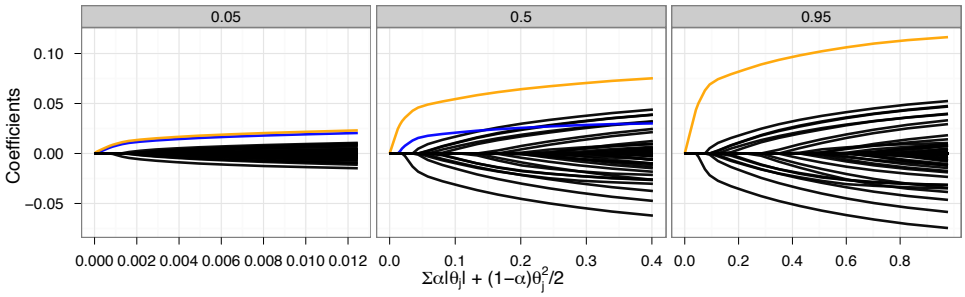
Figures 9(a) and 9(b) summarize the variable selection results for the logistic L_2E and MLE for $\alpha = 0.05, 0.5$, and 0.95. Recall from (5.4) that small α emphasizes the ridge part of the penalty while large α emphasizes the LASSO part of the penalty.

SNP markers can have a high degree of collinearity due to recombination mechanics. SNPs that are physically close to each other tend to be highly correlated and are said to be in linkage disequilibrium. The pair rs1051730 and rs8034191 for example are in "high" linkage disequilibrium.

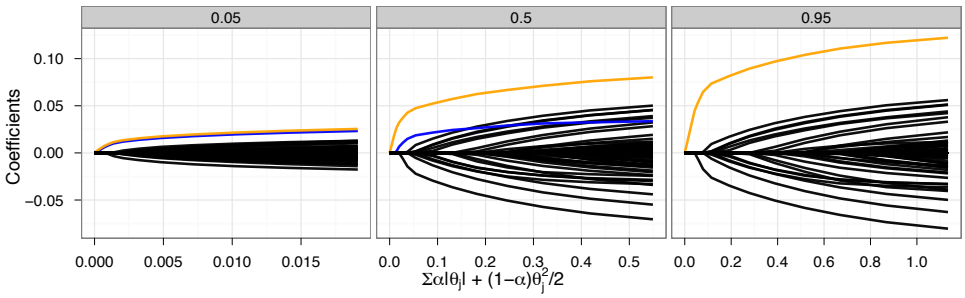
There are three things to note. First, the regularization paths for the L_2E and MLE are almost identical. Second, both methods produce regularization paths that identify rs1051730 (orange) and rs8034191 (blue) as having the greatest partial correlation the case/control status. Third, the paths for rs1051730 and rs8034191 behave as would be expected with α. For small α, or more ridge-like penalty, the two paths become more similar. For large α, or more LASSO-like penalty, only one of the two correlated predictors enters the model while the other is excluded.

9. Discussion. Outliers can introduce bias in some commonly used maximum likelihood estimation procedures. This well known fact, however, warrants attention because bias can have material effects on the ubiquitous LASSO-based variable selection procedures. In the context of standard logistic regression, influential outliers cause implosion breakdown. In this paper we have demonstrated that the combination of implosion breakdown and the soft-thresholding mechanism of LASSO variable selection can result in missed detection of relevant predictors.

To guard against the undue influence of outliers on estimation and variable selection, we propose a robust method for performing sparse logistic regression. Our method is based on minimizing the estimated L_2 distance between the logistic parametric model and the underlying true conditional distribution. The resulting optimization problem is a penalized non-linear least squares problem which we solve with an MM algorithm. Our MM



(a) L_2E



(b) MLE

FIG 9. *Regularization paths of regression coefficients of SNP markers on Chromosome 15 for L_2E and MLE for $\alpha = 0.05, 0.5,$ and 0.95. The regularization path for rs1051730 is in orange; the path for rs8034191 is in blue. The L_2E and MLE paths are nearly identical.*

algorithm in turn reduces the optimization problem into solving a series penalized least squares problems whose solution paths can be solved very efficiently with coordinate descent and warm starts.

Although we present our work as a method for robust binary logistic regression, our method immediately extends to other related contexts. Our algorithm can be extended to handle more than two classes. The generalization to the K-class multinomial is straightforward.

$$L(\mathbf{Y}, \tilde{\mathbf{X}}\boldsymbol{\Theta}) = \sum_{k=1}^{K} \|\mathbf{y}_k - F_k(\tilde{\mathbf{X}}\boldsymbol{\Theta})\|_2^2,$$

where $y_{ik} = 1$ if the ith observation belongs to class k and 0 otherwise and the ith element of vector $F_k(\tilde{\mathbf{X}}\boldsymbol{\Theta})$ is given by

$$\frac{\exp(\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\theta}_k)}{1 + \sum_{j=1}^{K} \exp(\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\theta}_j)}.$$

This non-linear least squares problem also has bounded curvature and consequently can also be solved by minimizing a sequence of LASSO-penalized least squares problems.

Our algorithm can also be used as a subroutine in performing robust binary principal component analysis and, more generally, robust binary tensor decompositions. A common strategy in array decompositions for multiway data, including multiway binary data, is to use block coordinate descent or alternating minimization (Chi and Kolda, 2011; Collins, Dasgupta and Schapire, 2002; Kolda and Bader, 2009; Lee, Huang and Hu, 2010). For binary multiway data, each block minimization would perform a robust logistic regression step.

We want to make clear that the logistic L_2E is not a competitor to the MLE but rather a complement. Both methods are computationally feasible and can be run on data together. As seen in the real data examples of Section 8, sometimes the logistic L_2E recovers the MLE solution. On the other hand, when discrepancies do occur, taking the MLE and L_2E solutions together can provide insight into the data that would be harder to identify with the MLE solution alone.

We close with some interesting directions for future work. We have seen that LASSO-based variable selection in the presence of implosion breakdown can lead to missed detection of relevant predictors. This motivates the question of whether explosion breakdown can lead to the inclusion of irrelevant predictors. Finally, with respect to convergence issues of our algorithm, while we have established conditions under which our algorithm is guaranteed to

converge to a stationary point we do not have rigorous results on the rate at which it does so. As a complement to methods that may be sensitive to the presence of outliers, characterizing the convergence speed of our algorithm has a great deal of practical importance.

The code for both logistic L_2E and HHSVM will be made available, and the authors hope to receive feedback on its use.

APPENDIX A: PROOFS

A.1. Proof of Theorem 5.1. It is immediate that $L(\tilde{\boldsymbol{\theta}}; \tilde{\boldsymbol{\theta}}) = L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})$. We turn our attention to proving that $L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) \geq L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbb{R}^{p+1}$. Since $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ has bounded curvature our strategy is to represent $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ by its exact second order Taylor expansion about $\tilde{\boldsymbol{\theta}}$ and then find a tight uniform bound over the quadratic term in the expansion. This approach applies in general to functions with continuous second derivative and bounded curvature (Böhning and Lindsay, 1988).

The exact second order Taylor expansion of $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$ is given by

$$L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) = L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}} \nabla L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}} \mathbf{H}_{\boldsymbol{\theta}^*} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where $\boldsymbol{\theta}^* = \gamma\tilde{\boldsymbol{\theta}} + (1-\gamma)\boldsymbol{\theta}$ for some $\gamma \in (0,1)$ and

$$\begin{aligned}
\nabla L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) &= 4n^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{G}(\mathbf{p} - \mathbf{y}) \\
\mathbf{H}_{\boldsymbol{\theta}} &= \frac{2}{n}\mathbf{X}^{\mathsf{T}}\mathbf{M}_{\boldsymbol{\theta}}\mathbf{X}, \\
\mathbf{G} &= \operatorname{diag}\{p_1, \ldots, p_n\} \\
\mathbf{M}_{\boldsymbol{\theta}} &= \operatorname{diag}\{\psi_{u_1}(p_1), \ldots, \psi_{u_n}(p_n)\} \\
\mathbf{u} &= 2\mathbf{y} - \mathbf{1} \\
\mathbf{p} &= F(\tilde{\mathbf{X}}\boldsymbol{\theta}) \\
\psi_u(p) &= [2p(1-p) - (2p-1)((2p-1) - u)]p(1-p).
\end{aligned}$$

Note that $(\mathbf{M}_{\boldsymbol{\theta}})_{ii}$ is bounded from above, i.e., $\sup_{\boldsymbol{\theta}\in\Theta}(\mathbf{M}_{\boldsymbol{\theta}})_{ii} < \infty$. We now introduce a surrogate function:

$$L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) + \frac{4}{n}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{G}(F(\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) - \mathbf{y}) + \frac{\eta}{n}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where

$$\eta \geq \max\left\{ \sup_{p\in[0,1]} \psi_{-1}(p), \sup_{p\in[0,1]} \psi_1(p) \right\}.$$

Note that for any $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$, $(\mathbf{M}_{\boldsymbol{\theta}})_{ii} \leq \eta$. Therefore,

$$\begin{aligned}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}} \mathbf{H}_{\boldsymbol{\theta}^*} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) &= (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{M}_{\boldsymbol{\theta}^*} \mathbf{X} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \\ &\leq \eta (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),\end{aligned}$$

and consequently $L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$ majorizes $L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})$ at $\tilde{\boldsymbol{\theta}}$. □

The following observations lead to a simpler lower bound on η. Note that

$$\sup_{p \in [0,1]} \psi_{-1}(p) = \sup_{p \in [0,1]} \psi_1(p),$$

since $\psi_{-1}(p) = \psi_1(1-p)$. So, the lower bound on η can be more simply expressed as

$$\sup_{p \in [0,1]} \psi_1(p) = \max_{p \in [0,1]} \psi_1(p) = \frac{1}{4} \max_{q \in [-1,1]} \left\{ \frac{3}{2} q^4 - q^3 - 2q^2 + q + \frac{1}{2} \right\}. \tag{A.1}$$

The first equality follows from the compactness of $[0,1]$ and the continuity of $\psi_1(p)$. The second equality follows from reparameterizing $\psi_1(p)$ in terms of $q = 2p - 1$. Since the derivative of the polynomial in (A.1) has a root at 1, it is straightforward to argue that the lower bound of η is attained at the second largest root, which is $(-3 + \sqrt{33})/12$. Thus, the majorization holds so long as

$$\eta \geq \frac{3}{16} q^4 - \frac{1}{4} q^3 - \frac{1}{2} q^2 + \frac{1}{4} q + \frac{1}{16} \Bigg|_{q = \frac{-3+\sqrt{33}}{12}}.$$

A.2. Proof of Theorem 5.2. A key condition in MM algorithm convergence proofs is coerciveness since it is a sufficient condition to ensure the existence of a global minimum. Recall that a continuous function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is coercive if all its level sets $S_t = \{\mathbf{x} \in U : f(\mathbf{x}) \leq t\}$ are compact.

We will use the MM algorithm global convergence results in Schifano, Strawderman and Wells (2010). Let $\xi(\boldsymbol{\theta})$ denote the objective function and let $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ denote a surrogate objective function that will be minimized with respect to its first argument in lieu of $\xi(\boldsymbol{\theta})$. The iteration map φ is given by

$$\varphi(\tilde{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\theta}} \xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}).$$

We now state a slightly less general set of regularity conditions than those in Schifano, Strawderman and Wells (2010) that are sufficient for our purposes. Suppose $\xi, \xi^{[S]}$, and φ satisfy the following set of conditions:

R1. The objective function $\xi(\boldsymbol{\theta})$ is locally Lipschitz continuous for $\boldsymbol{\theta} \in \Theta$ and coercive. The set of stationary points $\mathcal{S}$ of $\xi(\boldsymbol{\theta})$ is a finite set, where the notion of a stationary point is defined as in Clarke (1983).
R2. $\xi(\boldsymbol{\theta}) = \xi^{[S]}(\boldsymbol{\theta}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$.
R3. $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) < \xi^{[S]}(\boldsymbol{\theta}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$ where $\boldsymbol{\theta} \neq \tilde{\boldsymbol{\theta}}$.
R4. $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is continuous for $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \in \Theta \times \Theta$ and locally Lipschitz in Θ.
R5. $\varphi(\boldsymbol{\theta})$ is a singleton set consisting of one bounded vector for $\boldsymbol{\theta} \in \Theta$.

Then $\{\boldsymbol{\theta}^{(n)}, n \geq 0\}$ converges to a fixed point of the iteration map φ. By Proposition A.8 in Schifano, Strawderman and Wells (2010) the fixed points of φ coincide with $\mathcal{S}$.

In our case we have the following objective and surrogate functions

$$\xi(\boldsymbol{\theta}) = \frac{1}{2n}\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right)$$
$$\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \frac{1}{2}L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right).$$

We check each regularity condition in turn.

R1. Since $\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2$ is bounded below and the penalty term is coercive, $\xi(\boldsymbol{\theta})$ is coercive. Recall that the gradient of the $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ is $(4/n)\mathbf{X}^\mathsf{T}\mathbf{G}(F(\tilde{\mathbf{X}}\boldsymbol{\theta}) - \mathbf{y})$. The norm of the gradient is bounded; specifically it is no greater than $2\sigma_1^2$ where σ_1 is the largest singular value of $\mathbf{X}$. Therefore, $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ is Lipschitz continuous and therefore locally Lipschitz continuous. Consequently, $\xi(\boldsymbol{\theta})$ is locally Lipschitz continuous. If the set of stationary points of $\xi(\boldsymbol{\theta})$ is finite, then R1 is met.

R2 and R3. Recall the majorization we are using is given by

$$L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\mathsf{T}\nabla L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) + \frac{\eta}{n}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where

$$\eta > \frac{1}{4}\max_{q \in [-1,1]}\left\{\frac{3}{2}q^4 - q^3 - 2q^2 + q + \frac{1}{2}\right\}.$$

To ensure that the majorization is strict we need the inequality to be strict. Thus, the curvature of the majorization exceeds the maximum curvature of $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ and the majorization is strict. R2 and R3 are met.

R4. The penalized majorization is the sum of continuous functions in $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \in \Theta \times \Theta$ and is consequently continuous. The penalized majorization as a function of its first argument is the sum of a positive definite quadratic function and the 1-norm function, both of which are locally Lipschitz continuous so their sum is locally Lipschitz continuous. R4 is met.

R5. If $\lambda(1-\alpha) > 0$ then $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is strictly convex in $\boldsymbol{\theta}$ and thus has at most one global minimizer. Since $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is also coercive in $\boldsymbol{\theta}$ it has at least one global minimizer. R5 is met.

Thus, Algorithm 1 will converge to a stationary point of $\xi(\boldsymbol{\theta})$, provided that there are only finitely many stationary points and the coordinate descent minimization of the Elastic Net penalized quadratic majorization is solved exactly. □

REMARK 1. *If ξ does not have finitely many stationary points, it can be shown that the limit points of the sequence of iterates are stationary points and that the set of limit points is connected (Chi, 2011; Schifano, Strawderman and Wells, 2010).*

REMARK 2. *The iterate update $\boldsymbol{\theta}^{(m+1)} = \varphi(\boldsymbol{\theta}^{(m)})$ can be accomplished by any means algorithmically so long as the global minimum of the majorization is found. Iterates of coordinate descent are guaranteed to converge to a global minimizer provided that the loss is differentiable and convex and the penalty is convex and separable (Tseng, 2001). Thus, applying coordinate descent on the Elastic Net penalized quadratic majorization will find the global minimum.*

REMARK 3. *Our definition of stationary points has to change because the objective functions of interest are locally Lipschitz continuous and therefore differentiable almost everywhere except on a set of Lebesgue measure zero. Clarke (1983) defines and proves properties of a generalized gradient for locally Lipschitz functions. Apart from pathological cases, when a function is convex the generalized gradient is the subdifferential. See Proposition 2.2.7 in Clarke (1983). When a function is differentiable the generalized gradient is the gradient. Thus as would be expected a point $\mathbf{x}$ is a stationary point of a locally Lipschitz function if the function's generalized gradient at $\mathbf{x}$ contains $\mathbf{0}$.*

ACKNOWLEDGMENTS

We thank Christopher Amos for generously allowing us to work with the lung cancer data set. All plots were made using the open source R package ggplot2 (Wickham, 2009).

REFERENCES

AMOS, C. I., WU, X., BRODERICK, P., GORLOV, I. P., GU, J., EISEN, T., DONG, Q., ZHANG, Q., GU, X., VIJAYAKRISHNAN, J., SULLIVAN, K., MATAKIDOU, A., WANG, Y.,

Mills, G., Doheny, K., Tsai, Y.-Y., Chen, W. V., Shete, S. a., Spitz, M. R. and Houlston, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics* **40** 616–622.

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549-559.

Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust Statistics, Data Analysis, and Computer Intensive Methods, Lecture Notes in Statistics* (H. Rieder, ed.) **109** 17–34. Springer-Verlag, New York.

Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics* **40** 641-663. 10.1007/BF00049423.

Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika* **92** 724-731.

Carroll, R. J. and Pederson, S. (1993). On Robustness in the Logistic Regression Model. *Journal of the Royal Statistical Society. Series B (Methodological)* **55** pp. 693-706.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998). Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing* **20** 33–61.

Chi, E. C. (2011). Parametric Classification and Variable Selection by the Minimum Integrated Squared Error Criterion PhD thesis, Rice University.

Chi, E. C. and Kolda, T. G. (2011). Making Tensor Factorizations Robust to Non-Gaussian Noise Technical Report No. SAND2011-1877, Sandia National Laboratories, Albuquerque, NM and Livermore, CA.

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis*. Wiley-Interscience.

Collins, M., Dasgupta, S. and Schapire, R. E. (2002). A generalization of principal component analysis to the exponential family. *Advances in neural information processing systems* **1** 617–624.

Copas, J. B. (1988). Binary Regression Models for Contaminated Data. *Journal of the Royal Statistical Society. Series B (Methodological)* **50** pp. 225-265.

Croux, C., Flandre, C. and Haesbroeck, G. (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters* **60** 377–386.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association* **90** pp. 1200-1224.

Donoho, D. L. and Liu, R. C. (1988). The "Automatic" Robustness of Minimum Distance Functionals. *Annals of Statistics* **16** 552–586.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32** 407-499.

El Ghaoui, L., Viallon, V. and Rabbani, T. (2010). Safe Feature Elimination in Sparse Supervised Learning Technical Report No. UCB/EECS-2010-126, EECS Department, University of California, Berkeley.

Frank, A. and Asuncion, A. (2010). UCI Machine Learning Repository.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1** 302-332.

Genkin, A., Lewis, D. D. and Madigan, D. (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics* **49** 291-304.

Hunter, D. R. and Lange, K. (2004). A Tutorial on MM Algorithms. *The American Statistician* **58** 30–38.

KIM, J. and SCOTT, C. (2009). Performance analysis for L_2 kernel classification. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)* (D. KOLLER, D. SCHUURMANS, Y. BENGIO and L. BOTTOU, eds.) 833–840.

KIM, J. and SCOTT, C. (2010). L_2 Kernel Classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **32** 1822 -1831.

KOLDA, T. G. and BADER, B. W. (2009). Tensor Decompositions and Applications. *SIAM Review* **51** 455–500.

KÜNSCH, H. R., STEFANSKI, L. A. and CARROLL, R. J. (1989). Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models. *Journal of the American Statistical Association* **84** pp. 460-466.

LANGE, K. (2010). *Numerical Analysis for Statisticians*. Springer.

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization Transfer Using Surrogate Objective Functions. *Journal of Computational and Graphical Statistics* **9** pp. 1-20.

LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Annals of Applied Statistics* **4** 1579–1601.

LI, Y., DING, J. and ABECASIS, G. R. (2006). Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics* **S79** 2290.

LIU, Z., JIANG, F., TIAN, G., WANG, S., SATO, F., MELTZER, S. J. and TAN, M. (2007). Sparse Logistic Regression with Lp Penalty for Biomarker Identification. *Statistical Applications in Genetics and Molecular Biology* **6**.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, Boca Raton, Florida.

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Computational Statistics & Data Analysis* **52** 374 - 393.

NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical optimization. Springer series in operations research*. Springer.

OWEN, A. B. (2006). A robust hybrid of lasso and ridge regression Technical Report, Stanford University.

ROSSET, S. and ZHU, J. (2007). Piecewise linear regularized solution paths. *Annals of Statistics* **35** 1012–1030.

SCHIFANO, E. D., STRAWDERMAN, R. L. and WELLS, M. T. (2010). Majorization-Minimization algorithms for nonsmoothly penalized objective functions. *Electronic Journal of Statistics* **4** 1258–1299.

SCOTT, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*. John Wiley & Sons, Inc.

SCOTT, D. W. (2001). Parametric Statistical Modeling by Minimum Integrated Square Error. *Technometrics* **43** 274-285.

SCOTT, D. W. (2004). Partial mixture estimation and outlier detection in data and regression. In *Theory and Applications of Recent Robust Methods* (M. HUBERT, G. PISON, A. STRUYF and S. V. AELST, eds.) 297-306. Birkhauser, Basel.

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** pp. 267-288.

TIBSHIRANI, R., BIEN, J., FRIEDMAN, J., HASTIE, T., SIMON, N., TAYLOR, J. and TIBSHIRANI, R. J. (2011). Strong rules for discarding predictors in lasso-type problems. arXiv:1011.2234v2.

TSENG, P. (2001). Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *Journal of Optimization Theory and Applications* **109** 475-494. 10.1023/A:1017501703105.

WANG, L., ZHU, J. and ZOU, H. (2008). Hybrid huberized support vector machines for

microarray classification and gene selection. *Bioinformatics* **24** 412-419.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics* **2** 224-244.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. and Lange, K. (2009). Genomewide Association Analysis by Lasso Penalized Logistic Regression. *Bioinformatics* **25** 714-721.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320.

Department of Human Genetics
University of California
Los Angeles, California, 90095, USA.
E-mail: ecchi@ucla.edu

Department of Statistics
Rice University
Houston, TX 77005, USA.
E-mail: scottdw@rice.edu